Computational Theory of Mind for Human-Agent Collaboration

Emre Erdogan[1][0000−0002−2139−3750], Frank Dignum[1,2][0000−0002−5103−8127], Rineke Verbrugge[3][0000−0003−3829−0106], and Pınar Yolum[1][0000−0001−7848−1834]

[1] Utrecht University, Utrecht, Netherlands
{e.erdogan1,p.yolum}@uu.nl
[2] Umeå University, Umeå, Sweden
dignum@cs.umu.se
[3] University of Groningen, Groningen, Netherlands
l.c.verbrugge@rug.nl

Theory of Mind (ToM) refers to the human ability of reasoning about the mental content of other people such as their beliefs, desires, and goals [4, 2]. This capability enables a human to reason about others, making it possible to understand, explain, and predict their behaviour. Being an important part of social cognition, having a functional ToM is especially useful when people collaborate, since a person can then effectively reason on what the other person knows as well as what reasoning they might do.

An important area where ToM could be of particular use is hybrid intelligence [1], where an agent can collaborate with a human towards a particular goal. For a successful human-agent collaboration to take place, an agent should be able to understand the human's behaviour, communicate well with the human, and continuously learn from their shared experience as well. We argue that such an agent would benefit from having a functional ToM for the human in achieving their collective goal in such hybrid settings.

In this paper, we summarize our work in which we propose an abstraction framework for computational modeling of ToM reasoning [3]. The main idea is to use an agent's set of beliefs and knowledge to produce more abstract, complex concepts for the agent to benefit from when interacting with humans. These complex concepts can correspond to various social norms, roles, as well as human values. Collectively, they serve as human-inspired heuristics, which can help the agent and the human to reach decisions.

To investigate the principle of abstraction, we provide an example hybrid setting in which an agent doctor and a human doctor collaborate towards a medical diagnosis of a patient's health problem [3]. Figure 1 outlines the interaction that takes place among the agent doctor A, the human doctor B, and the patient C during the diagnostic process. To make the setting more concrete, we computationally model several human decision-making heuristics and show how abstracting beliefs and knowledge into higher-level concepts can be useful for an agent doctor when doing decision-making with a human doctor. We emphasize how social dynamics shaped by roles, norms, and human values can play important parts in such hybrid settings. Our detailed examples demonstrate how such social dynamics can facilitate decision-making. We briefly sketch several ways

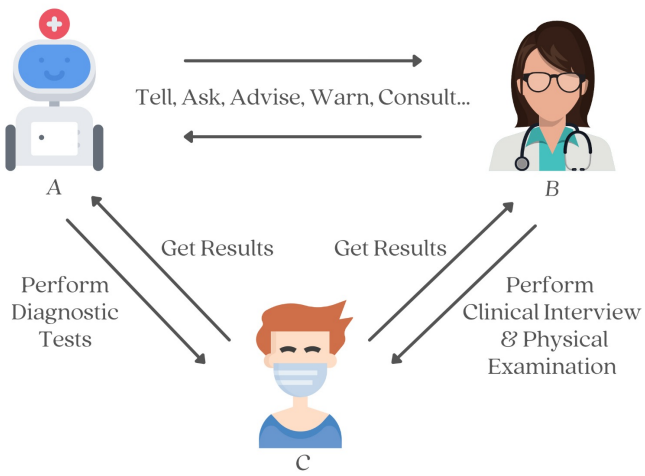

Fig. 1: Hybrid Collaboration in Medicine: An agent doctor (A) and a human doctor (B) work together towards the diagnosis of a patient's (C) health problem. Each doctor has different set of capabilities that would be useful for the diagnosis.

that the agent can employ these social dynamics within its belief abstraction and decision-making processes to achieve effective human-agent collaboration.

The idea of employing abstraction for computational ToM provides a base to build upon. As a follow-up work, we aim for a more concrete abstraction model in which we formalize the entities in our abstraction framework (i.e., beliefs, abstractions, etc.). Furthermore, we plan to benefit from various theories and methods in cognitive sciences since we aim for designing social agents that are capable of doing interactive recursive reasoning to collaborate efficiently with humans. With a more comprehensive agent model, which is also equipped with mind abstraction abilities, we will further test our agents in human-agent settings to evaluate their collaborative skills in dynamic environments.

References

1. Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al.: A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(08), 18–28 (2020)
2. Carruthers, P., Smith, P.K.: Theories of theories of mind. Cambridge University Press (1996)
3. Erdogan, E., Dignum, F., Verbrugge, R., Yolum, P.: Abstracting minds: Computational theory of mind for human-agent collaboration (2022), paper presented at the International Conference on Hybrid Human-Artificial Intelligence, VU Amsterdam.
4. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4), 515–526 (1978)